A t-test to compare means of two populations

- Assumptions:
  - Sample points are iid.
  - Distributions are approximately normal
  - Variances are approximately equal (homogeniety of variance).
- The t-test can be a paired test when the two samples come from the same population but treatment for each is different or it can be a two-sample t-test when the samples are from different populations.
  For two samples the t-statistic is calculated by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_p^2(\frac{1}{N_1} + \frac{1}{N_2}))}}$$

  where s_p^2 is the pooled variance given by: $s_p^2 = \frac{\sum_{i=1}^{k}(N_i - 1)s_i^2}{\sum_{i=1}^{k}(N_i - 1)}$, $s_i^2 = \frac{\sum_{j=1}^{N_i}(x_j - \bar{x}_i)^2}{N_i - 1}$. In this case $k = 2$.

Can we use multiple t-tests instead of ANoVA?

- No. Multiple t-tests instead of ANoVA will be incorrect.
- We know $P(\text{Type 1 error}) = \alpha$. This implies that the probability of no type 1 error for a single hypothesis test is $(1 - \alpha)$. If there are k such independent tests then the probability that there is no such type 1 error in all k tests is $(1 - \alpha)^k$. The probability that there will be at least one type 1 error in k tests, purely statistically, is:
  $P(\text{At least 1 type 1 error}) = (1 - (1 - \alpha)^k)$.
- Assume $\alpha = 0.05$ and $k = 10$.
  $P(\text{At least one type 1 error}) = (1 - 0.95^{10}) \approx 0.4$. So, there is a substantial probability of a type 1 error happening purely by chance. This is the general problem of multiple hypothesis testing. So, one should use ANoVA instead of multiple t-tests.

Corrections for multiple hypothesis tests

- In the event multiple hypotheses have to be tested then one has to use corrections to ensure that the overall type 1 error rate is bounded by α.
- One widely used correction is Bonferroni's correction. So, if k is the number of hypotheses to be tested then the α level for each individual test is set to $\frac{\alpha}{k}$.
- Bonferroni's correction happens to be a conservative correction and the overall level generally is less than α. There are other measures and corrections that have been proposed.
- Reducing α for the individual tests increases β and thereby reduces power.

Correlation, Regression

- Hypothesis testing is typically needed when we have a hypothesis about population parameters. We have seen the related theory and many examples.
- A research question typically wants to understand the relationship between independent variables that are being manipulated and the dependent variable(s). That is we want a predictive relationship between the independent and dependent variable(s).
- Two tools are:
  - Correlational analysis (qualitative).
  - Regression analysis (quantitative).

Correlation

- Two variables X and Y are correlated if they have something in common. So, for pairs (X, Y) one simultaneously observes either higher values of X and Y (positive correlation) or higher values of X and lower values of Y (negative correlation). Both X, Y are numeric variables - most often interval or ratio variables.
- The intensity of the correlation is measured by the correlation coefficient (called Pearson correlation coefficient), symbolized by r, which ranges between $+1$ and -1. The assumption is there is some underlying factor(s) (often called latent variable(s)) that connects the two variables. r is the correlation coefficient for the sample. The population coefficient is denoted by ρ.
- This is strictly **linear** correlation. If X, Y vary non-linearly then r is not meaningful (see later for an example).

Calculation of ρ and r

Given RVs X, Y the population Pearson correlation coefficient is:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_x)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

For sample of size N, sample values are used. The sample Pearson correlation coefficient is:

$$r_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

$\bar{x}$ and $\bar{y}$ are the sample means for X and Y respectively.

Examples of positive and negative correlation

- Examples of positive correlation:
  - Height and weight of an individual.
  - Years of education and earning per year at age 50.
  - Husband's age and wife's age in a married couple.
  - JEE Adv. rank and graduating CPI in the BTech/BS program at IITK.
- Examples of negative correlation:
  - Obesity and life-span of an individual.
  - Height above sea-level and atmospheric air pressure on earth.
  - Average day-time temperature and latitude North or South on earth.
  - Length of a list of words and speed of recall of a target word in the list.

Concrete correlation example

Word	Length	Number of Lines
bag	3	14
across	6	7
on	2	11
insane	6	9
by	2	9
monastery	9	4
relief	6	8
slope	5	11
scoundrel	9	5
with	4	8
neither	7	2
pretentious	11	4
solid	5	12
this	4	9
for	3	8
therefore	9	1
generality	10	4
arise	5	13
blot	4	15
infectious	10	6
$\sum$	120	160
M	6	8

Figure: Data on length of words in characters and length of definition in lines in the OED. (From Abdi etal, Experimental Design and Analysis for Psychology.)

Correlation example: scatter plot

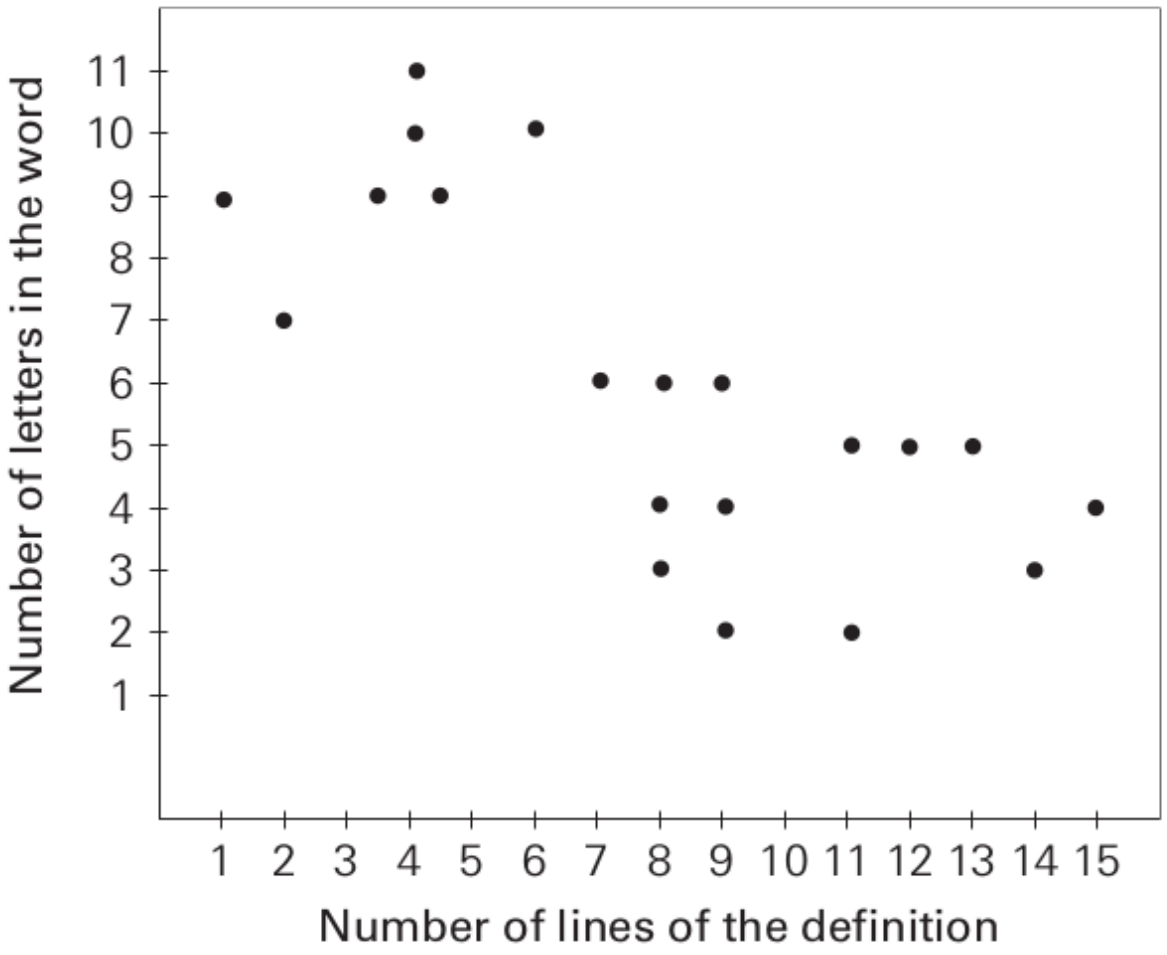


Figure: Scatter plot of the word-length vs definition-length. (From Abdi etal, Experimental Design and Analysis for Psychology.)

Correlation example: mean shifted to 0

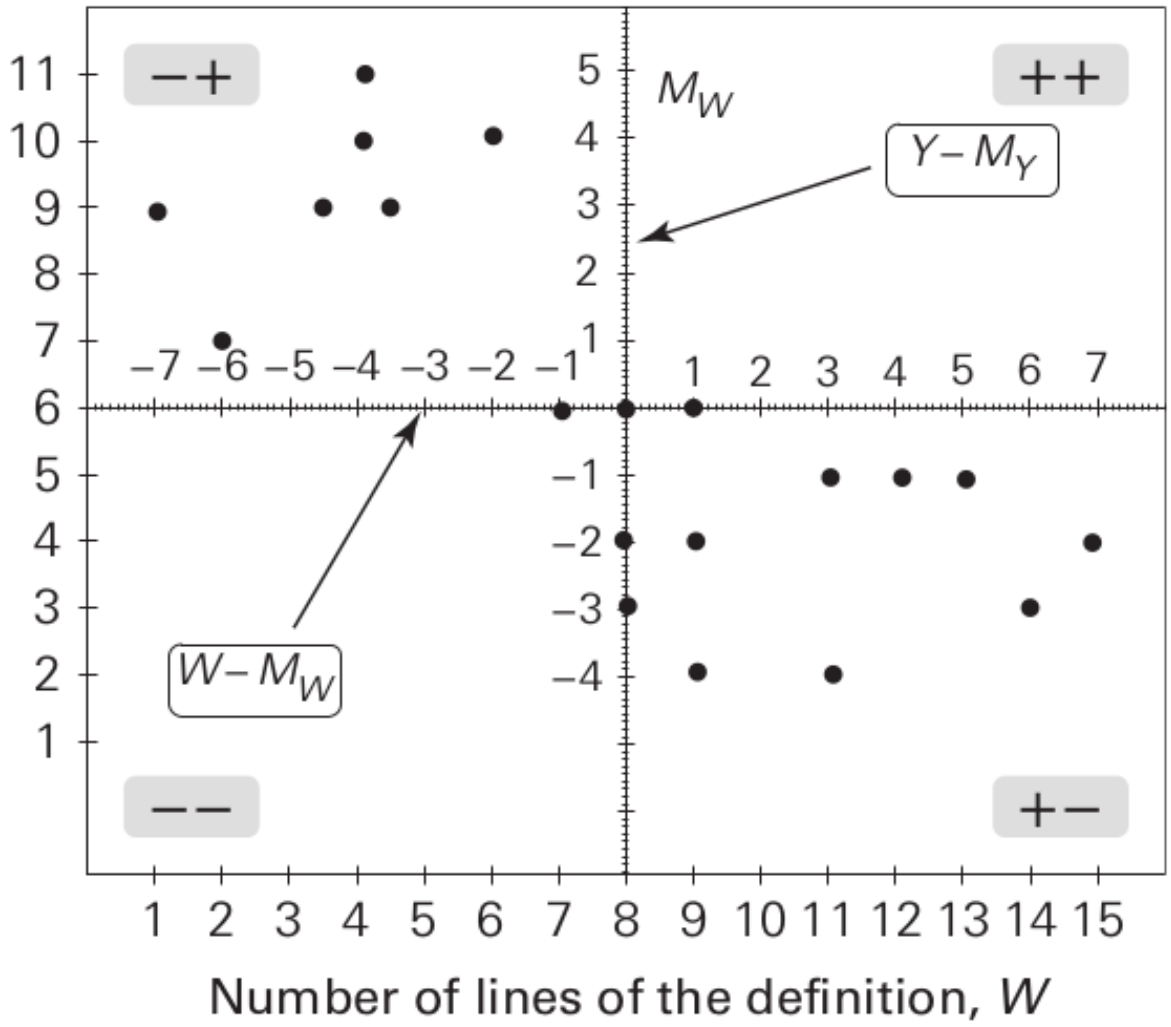


Figure: The previous figure after mean shifted to 0. (From Abdi etal, Experimental Design and Analysis for Psychology.)

Effect of outliers

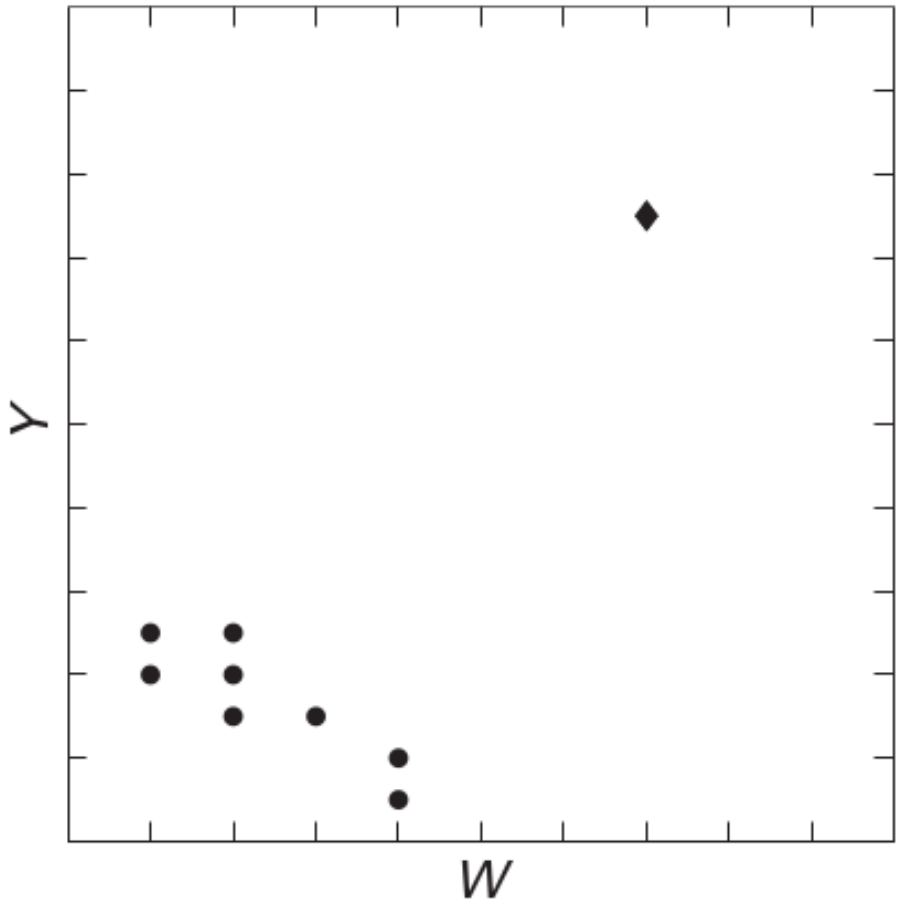


Figure: Correlation coefficient r=-0.87 without the diamond point. With the diamond point it is r=0.61. (From Abdi etal, Experimental Design and Analysis for Psychology.)

Correlation vs Causation

- **Correlation does not imply causation**. We have seen this earlier. Here is a factoid:
  In tier 1 and tier 2 cities in India the number of crimes is highly correlated with the number of schools.
  We cannot conclude schools lead to crimes.
- However, if two dependent variables are correlated then one can be used to predict the other. For example, an index derived from presence of some domestic appliances (e.g. TV, telephone, fridge) and some kinds of books in a household showed a correlation of 0.8 for verbal intelligence of children in the household.
- Often correlations amongst dependent variables is studied to find an underlying abstract or latent factor that can explain the correlation.
- Correlation between dependent variables can be used as an experimental first step before looking for causation. Example: imagery affects memory. Check correlation between highly visual persons (those who use imagery spontaneously) and their memory.

F-ratio

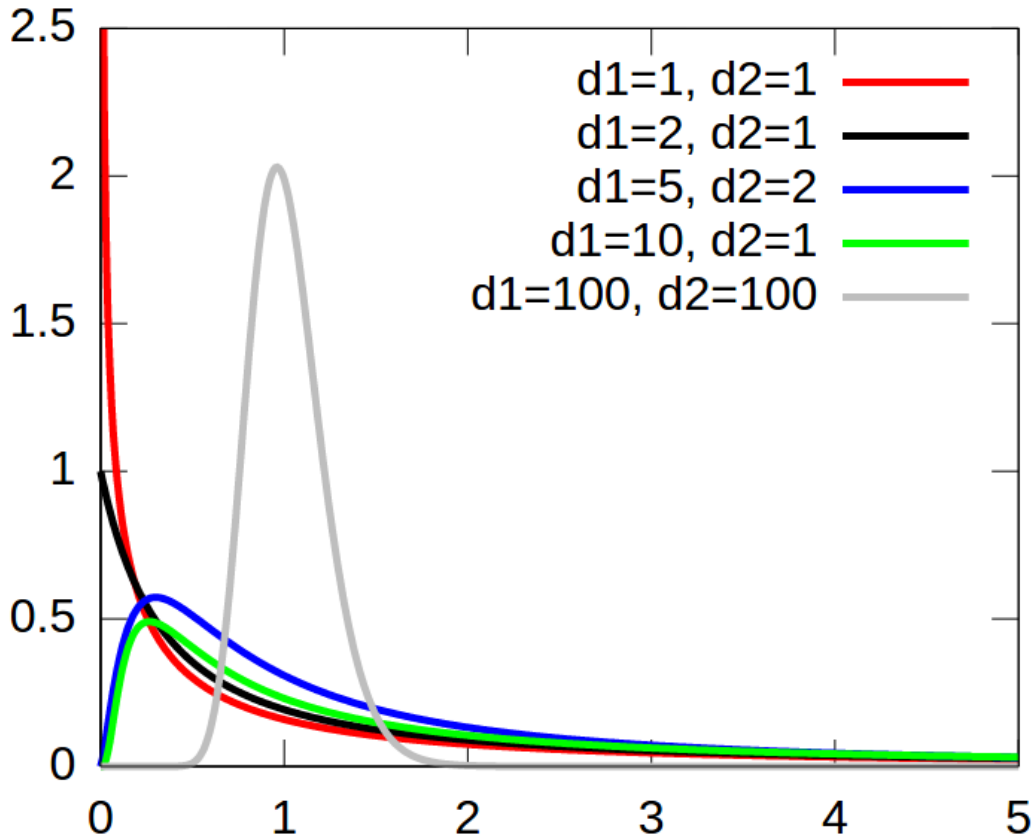
- Given a value of r for a sample how do we know that it holds for the population?
- Again we have to use hypothesis testing. The null hypothesis is $H_0 : \rho = 0$ (ρ is correlation coefficient for the population). That is there is no correlation between the variables in the population.
- The statistic used is the F-ratio (Fisher ratio).

$$F = \frac{r^2}{1 - r^2} \times (N - 2)$$

$N - 2$ is called the *degrees of freedom*. N is the sample size.

The F distribution

The F distribution (Fisher) has two parameters ν_1 and ν_2 (the degrees of freedom). The distribution can be thought of as a ratio of two χ^2 variates from where we get ν_1 and ν_2. It is also the sampling distribution of the F-ratio when the null hypothesis is true. So, it can be used to get critical values.



Source: Wikipedia

The test

- The sampling distribution F has $\nu_1 = 1$ called the correlation degrees of freedom and $\nu_2 = N - 2$. ν_1 is always 1 for correlations. Requires that both distributions for X, Y are normal.
- The critical value can be found from tables of the F distribution that are parametrized by ν_1, ν_2. See figure below for an example. Value of statistic should be $\geq$ critical value to reject H_0.

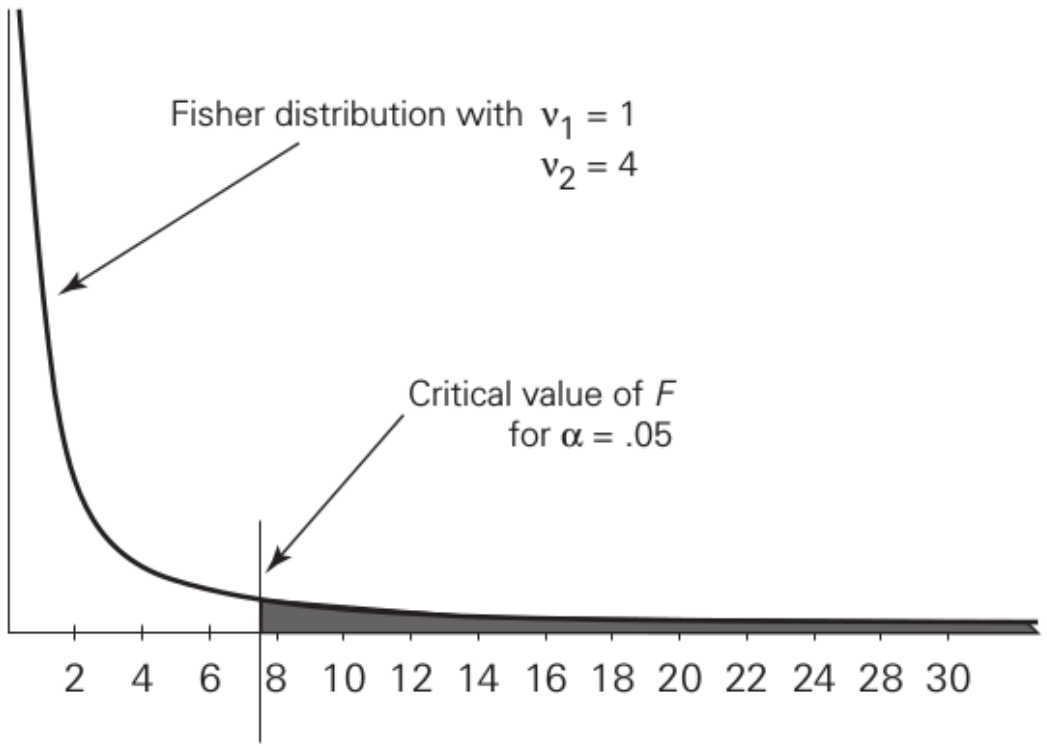


Table fragment for $F_{critical}$

Table 2 Critical values of Fisher's *F*. $\alpha = .05$ $\alpha = .01$

	ν_1												
ν_2	1	2	3	4	5	6	7	8	9	10	11	12	14
1	161 **4052**	199 **4999**	215 **5403**	224 **5624**	230 **5763**	233 **5858**	236 **5928**	238 **5981**	240 **6022**	241 **6055**	242 **6083**	243 **6106**	245 **6142**
2	18.51 **98.50**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.35 **99.36**	19.37 **99.37**	19.38 **99.39**	19.40 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**
3	10.13 **34.12**	9.55 **30.82**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.89 **27.67**	8.85 **27.49**	8.81 **27.35**	8.79 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.55**	5.94 **14.45**	5.91 **14.37**	5.87 **14.25**
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.46**	4.82 **10.29**	4.77 **10.16**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**
6	5.99 **13.75**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **6.99**	3.73 **6.84**	3.68 **6.72**	3.64 **6.62**	3.60 **6.54**	3.57 **6.47**	3.53 **6.36**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.18**	3.44 **6.03**	3.39 **5.91**	3.35 **5.81**	3.31 **5.73**	3.28 **5.67**	3.24 **5.56**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.61**	3.23 **5.47**	3.18 **5.35**	3.14 **5.26**	3.10 **5.18**	3.07 **5.11**	3.03 **5.01**

Example

We will use the following sample data:

(x_1, y_1)	(x_2, y_2)	(x_3, y_3)	(x_4, y_4)	(x_5, y_5)	(x_6, y_6)
$(1, 16)$	$(3, 10)$	$(4, 12)$	$(4, 4)$	$(5, 8)$	$(7, 10)$

We get:

$\mu_X = 4,\ \mu_Y = 10$

$SCP_{XY} = -20,\ SS_X = 20,\ SS_Y = 80$

$r = \frac{SCP_{XY}}{\sqrt{SS_X SS_Y}} = \frac{-20}{\sqrt{20 \times 80}} = -0.5$

$F = \frac{r^2}{1-r^2}(N-2) = \frac{0.25}{0.75}(6-2) = 1.33$

Assuming $\alpha = 0.05$ we get $F_{critical} = 7.71$ (see table in previous slide).
Since $F < F_{critical}$ the null hypothesis H_0 cannot be rejected.

Correcting r for the population

- The r obtained from the sample is a biased estimate for the population. It is an over-estimate. So, we need a correction.
- This is similar to σ where the sample std. deviation s is biased and is lower than the population standard deviation σ.
- The correction for r is much more complex and there are multiple correction formulae available. Below are two such formulae. The first is the one most often used in statistical s/w packages. The second (Stein's correction) is more accurate but also more complex. Both formulae use r^2 instead of r so we should use the correct sign when we take the sqrt. We will represent the corrected coefficient by r'.
  1. $r'^2 = 1 - [(1 - r^2)(\frac{N-1}{N-2})]$
  2. $r'^2 = 1-[(1-r^2)(\frac{N-1}{N-2})(\frac{N-2}{N-3})(\frac{N+1}{N})] = 1-[(1-r^2)(\frac{N-1}{N-3})(\frac{N+1}{N})$

Correction for the example

Below we calculate the corrected r' using both formulae. Recall that $r = -0.5$ in the example.

1. $r'^2 = 1 - [(1 - 0.25)\frac{5}{4}] = \frac{1}{16}$. So, $r' = -0.25$. Corrected value is half of the sample coefficient which was -0.5.
2. $r'^2 = 1 - [(1 - r^2)(\frac{5}{3})(\frac{7}{6}] = -\frac{11}{24}$. A negative value for r'^2 means we must conclude $r' = 0$. Stein's formula typically gives values smaller than the first formula.

Using empirical Monte Carlo method for testing

- The standard method for hypothesis testing makes some assumptions. For the correlation coefficient, Fisher assumed that both X and Y were normally distributed.
- Due to computers a widely applicable general method to test hypotheses by actual sampling is now available. The Monte Carlo method.
- H_0 true implies X, Y are uncorrelated and independent. So, sample 6 values from one normal distribution and another 6 from another independent normal distribution and pair them up. Then compute r^2 and the F-ratio. Repeat process a large number of times to get a histogram of F values - approx. F distribution. Calculate $F_{critical}$ the value of the F-ratio for the 95^{th} percentile corresponds to $\alpha = 0.05$ - an empirically determined critical value. The rule for rejecting H_0 remains the same sample $F \geq F_{critical}$.
- One advantage of this approach is that it can be made to work for any distribution for X, Y. Not constrained to just the normal distribution.